MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2019/2020

TMA 7021 – DATA MINING AND ANALYTICS

(TC01)

13 JANUARY 2020
3.00 p.m – 5.00 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of 7 pages with FOUR questions only.
2. Answer **ALL** questions.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

(a) The strength of a Bayesian Network is that it has both the *quantitative* and *qualitative* portions. Briefly discuss the two portions.

[2 marks]

(b) You are requested to design and draw a Bayesian Network for motor claim fraud detection. Fraud can be inferred from (i) suspicious driver characteristics, (ii) vehicle age, (iii) weather condition, (iv) road condition. Suspicious driver characteristics can be inferred from the driver's *age*, *gender* and *years of driving experience*.

[4 marks]

(c) Name the states (values) for the following nodes:

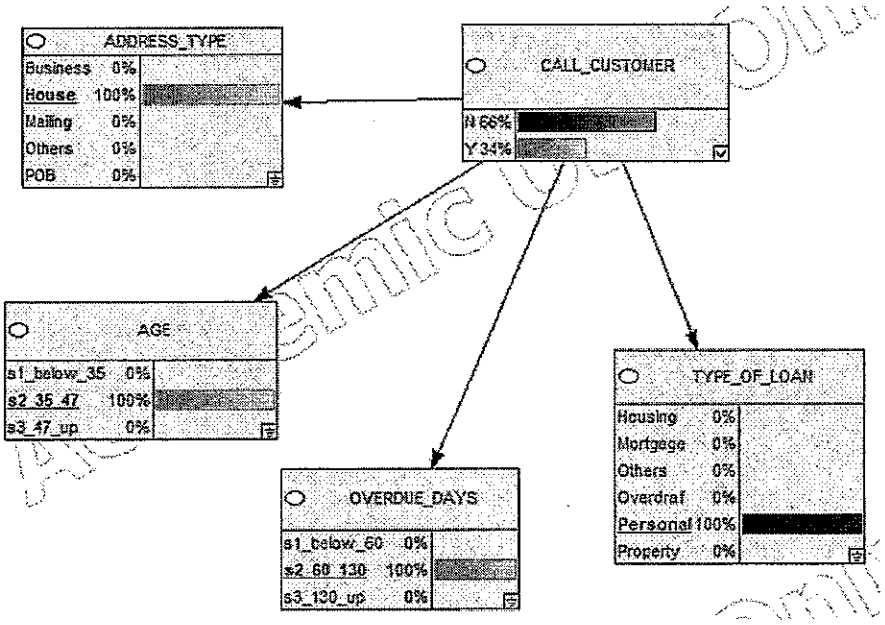
a. Driver characteristics

b. Years of driving experience

[2 marks]

(d) Write the corresponding rule after evidence instantiation on the Bayesian Network below:

[2 marks]



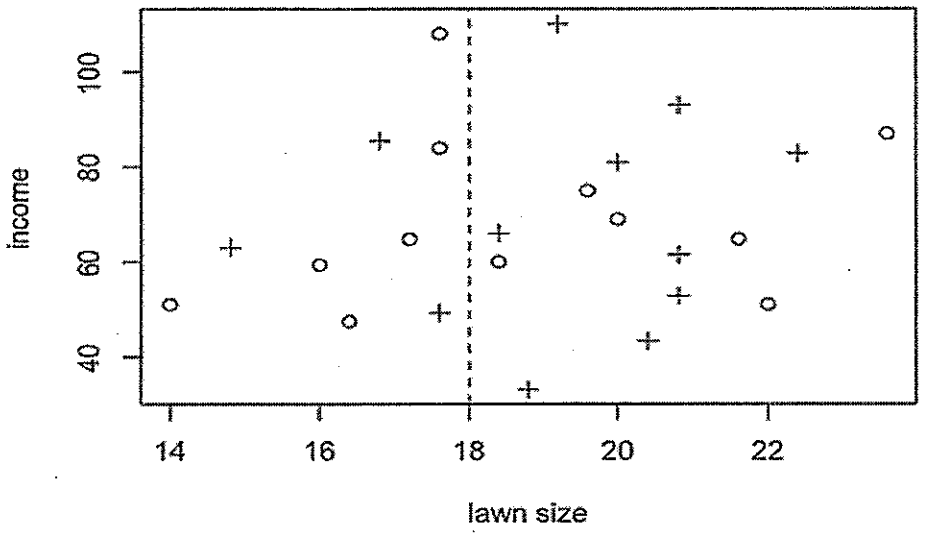
Continued...

QUESTION 2

(a) Name TWO (2) methods to measure node impurity in a decision tree.

[2 marks]

(b) Scatterplot below shows the relationship between house lawn size and household income. The "+" symbol represents *male* while the "o" symbol represents *female*.



(i) Calculate the Gini Index of the entire dataset before splitting the dataset at house size = 18.

[2 marks]

(ii) Calculate the reduction in Gini Index by introducing a split of dataset at lawn size = 18.

[2 marks]

(iii) Write one possible rule generated based on the diagram above.

[2 marks]

(c) There are three different types of feature selection methods, namely, *filter*, *wrapper*, *embedded*. Why is the *filter* method is commonly used as benchmark to *wrapper* and *embedded* methods?

[2 marks]

Continued...

QUESTION 3

(a) The table below shows the movies purchased by customers.

customer	Movies
1	Toy Story 4, Avengers, Spiderman
2	Spiderman, Black Panther
3	Logan, Black Panther
4	Toy Story 4, Avengers, Spiderman
5	Avengers, Toy Story 4

(i) Calculate the *support* for the following item-set:

{Logan, Black Panther}

[1 mark]

(ii) Calculate the *confidence* for the following rule:

{Spiderman, Avengers} → {Toy Story 4}

[1 mark]

(iii) Calculate the *Lift* for the following association rules:

{Toy Story 4, Avengers} → {Spiderman}

[1 mark]

(b) What can you conclude if Rule1 has a *lift* of 3.00 while Rule2 is 1.50?

[2 marks]

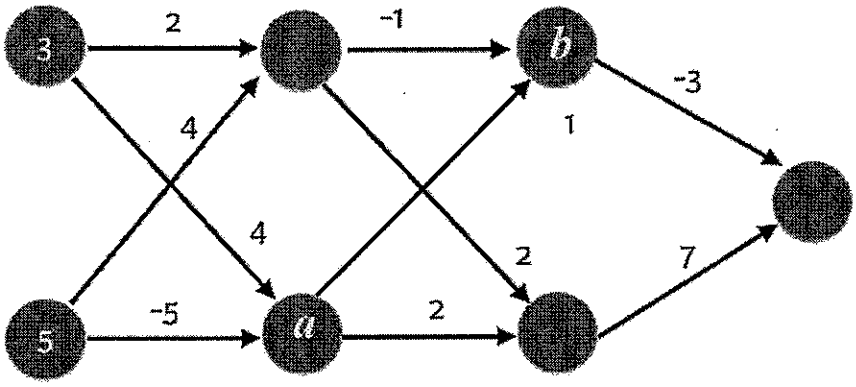
(c) Decision Tree is considered *supervised* machine learning while association rule mining is *unsupervised*. What is the main difference in term of rules generated by the two approaches?

[3 marks]

Continued...

(d) The figure below shows a simple neural network with one hidden layer. Suppose that each layer is calculated based on *ReLU*, calculate the value for a and b.

[2 marks]



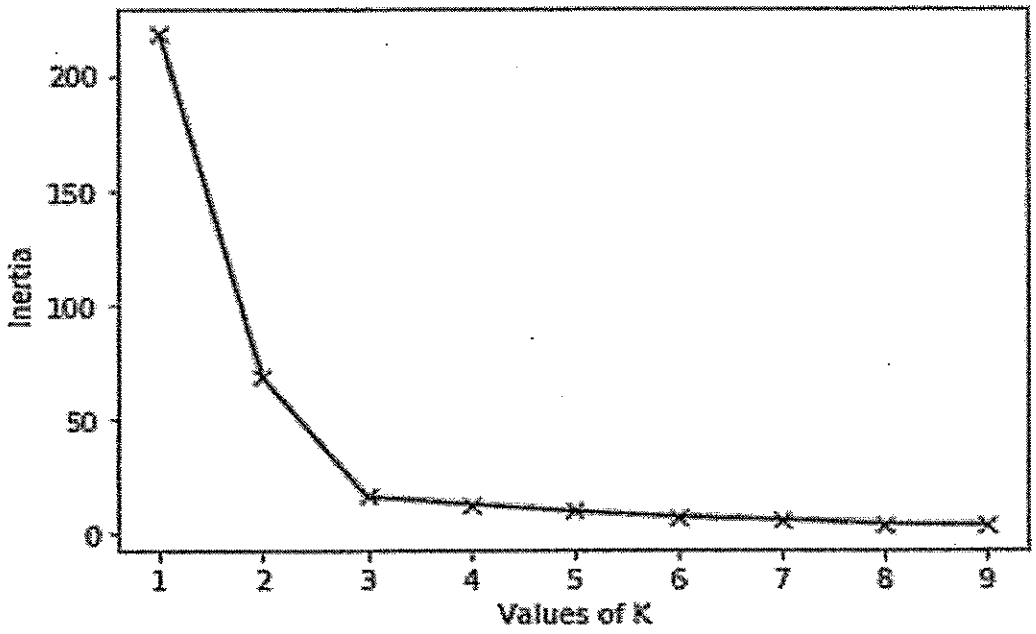
Continued...

QUESTION 4

(a) Briefly explain and differentiate between k-Means and k-NN.

[2 marks]

(b) The *elbow* method is used in k-means to determine the best number of clusters. From the diagram below, what is the k value? Explain how k is chosen.



[4 marks]

Continued...

(c) Non-Performing Loan (NPL) is something that a bank concerns a lot. It is commonly known as a loan that is in default or close to being in default. Study the table below which contains information about three customers of a bank.

Customer	Age	Credit_Score	Loan_Amount_in_K
Yen	25	3.2	89
Ting	18	3.1	98
LeLe	22	2.6	88

(i) Using *Euclidean* distance measurement technique, calculate the similarity for {Yen and LeLe} and {Yen and Ting}.

[2 marks]

(ii) Suppose we know that Yen's loan is tagged as NPL, based on your finding in Question 4(c)(i) above, who among *Ting* and *LeLe* is more likely to have NPL also? Justify your reason.

[2 marks]

End of Pages.